US House Price Analysis

25.11.2023

Priyanshu Sarkar

Github

Outline

- Executive Summary
- Goals
- Methodology
- Results
- Conclusion

Executive Summary

The Project involves conducting an analysis of various factors influencing home prices over the last two decades through a data science model. This endeavor aims to uncover the relationships between different variables and their impact on the fluctuations in housing prices.

The objective is to construct a robust data science model, likely employing regression analysis or more advanced machine learning techniques, to elucidate the correlations and predictive patterns among these factors and home prices. The model's output would ideally offer insights into the magnitude and direction of influence each variable holds over housing prices.

Goals

1. A robust predictive model that effectively estimates changes in the S&P/Case-Shiller U.S. National Home Price Index based on economic and housing-related factors.
2. Derive Factors that impact Price Index.

Methodology

Data Gathering (Link)

Most Important part before working on real time project is to about what you really looking for I don't know anything about House Price index so following questions I asked.

- Where can I learn about Price Index Link?
- What are the key factors that really affect price index Link?
- Where can we easlly get the relevant data Link?

In detail you can go and check how and where I downloaded and merge the data. I use the pandas python library for processing and merging different comma separated files.

Data Analysis (Link)

In the data analysis section first we try to understand the features which we gathered and processed in the Data Gathering section.

In which we ask basic analytical questions

- What are the data dimensions we are working with? (86, 8)
- What are the data types of each feature that help us to decide which test and visualization we have to work with?
- Do we have to deal with null or invalid observations? (No)
- How does data look mathematically like Mean, Median Standard Deviations, is the data skewed or Normal? (Sort of Normal)
- Do we have multicollinearity? (Yes)

Exploratory Data Analysis (EDA)

Univariate Analysis

- Two observations 2021-04-01 and 2021-07-01 are extreme values. After investigation, we discovered that this is due to high values in GDP and PERMIT, and lower values in INTDSRUSM193N. These factors led to higher values in CSUSHPISA.
- Higher values of Gross Domestic Product (GDP) lead to increased incomes for people, increased consumer spending, and can attract more investments due to perceived stability and profit potential. In 2021, the U.S. GDP was $23,315.08B, a 10.71% increase from 2020.
- Higher values of New Privately-Owned Housing Units Authorized in Permit-Issuing Places (PERMIT) lead to more housing availability and options for people to buy or rent.
- Lower values of INTDSRUSM193N (Interest Rates, Discount Rate for United States) lead to increased funds available for investments.
- These factors collectively contribute to very high values in CSUSHPISA. These factors should not be removed or treated as outliers.
- All features exhibit a sort of normal distribution, although a few are skewed. Upon investigation, these distributions are genuine, leading to the presence of outliers. It's important to include these features in the analysis.
- Consider trying different transformations to handle skewed features with outliers.
- Some independent columns demonstrate multicollinearity, affecting other independent features. However, these relationships are genuine and can be utilized in model building.

Multivariate Analysis

- Vacant Housing Units (EVACANTUSQ176N) - This feature shows a substantial correlation with CSUSHPISA, suggesting its influence on housing price variations.
- Gross Domestic Product (GDP) - GDP exhibits a significant connection with CSUSHPISA, indicating an impact on housing prices aligned with GDP fluctuations.
- Median Sales Price (MSPUS) - MSPUS demonstrates a robust correlation with CSUSHPISA, suggesting a direct influence on housing price trends.
- These features, Vacant Housing Units, Gross Domestic Product, and Median Sales Price, exhibit a noteworthy relationship with the target column CSUSHPISA, emphasizing their relevance in understanding housing price variations.

Final EDA Conclusion

- We had outliers in a few features but after transformation those extreme values were handled.
- No effect on correlation observed for features like EVACANTUSQ176N, GDP, MSACSR, and MSPUS after transformations.
- We can confidently apply Box Cox and StandardScaler.
- Overall, the distribution remains quite similar for all features before and after transformation.
- We have a multicollinearity problem (Either use VIF or Use them as it is.)

Model Development & Model Evaluations (Link)

In this section we try different base machine learning models and find the best performers. Once we find the top 2-3 models then tune the hyper parameters and compare their performance using performance metrics, Since we are working on a regression problem we use regression metrics.

Here are the models and their performance on train and test data. We use R2 Score for model performance but the core performance metric is Root Mean Squared Error because it is easy to interpret (same unit as target features units).

- Name: LinearRegression TrainR2: 0.97, TestR2: 0.95, DiffR2: 0.02
- Name: LinearRegression TrainRMSE: 0.08, TestRMSE: 0.09, DiffRMSE: -0.01

- Name: BayesianRidge TrainR2: 0.97, TestR2: 0.95, DiffR2: 0.02
- Name: BayesianRidge TrainRMSE: 0.08, TestRMSE: 0.09, DiffRMSE: -0.01

- Name: ElasticNet TrainR2: 0.0, TestR2: -0.07, DiffR2: 0.07
- Name: ElasticNet TrainRMSE: 0.49, TestRMSE: 0.44, DiffRMSE: 0.05

- Name: SGDRegressor TrainR2: 0.94, TestR2: 0.86, DiffR2: 0.08
- Name: SGDRegressor TrainRMSE: 0.12, TestRMSE: 0.16, DiffRMSE: -0.04

- Name: SVR TrainR2: 0.97, TestR2: 0.96, DiffR2: 0.01
- Name: SVR TrainRMSE: 0.09, TestRMSE: 0.09, DiffRMSE: -0.0

Question: Why not use tree-based or deep neural networks?

We do not want to draw decision boundaries because, as we saw while analyzing things like GDP, Median Sales, or Vacant, or any other aspect affects our Target columns.

The point here is in the real world, we can't say things go as we expected. That's why we want our model to learn from real-time data observations and predict outcomes based on that.

In tree-based models, we draw decision boundaries that are good for scenarios with limited possibilities.

Then what is wrong with the Neural Network? We want to know which factors affect more or less in predicting outcomes, which is not known if we work with neural networks.

Hence, we go with Linear Models. The advantage is they are less complex, and we say there is a linear correlation between the target and independent columns.

Question: What features contributes in predicting House Price Indexes (CSUSHPISA)

We are going to experiment with `Linear Regression` and `SVM` because they are the best performers.

Factors (Feature) that have high contributes in predicting target column are as follows

- GDP
- MSACSR
- PERMIT
- MSPUS

Conclusion

SVR model is clear better performer

Root Mean Squared Error : 0.0867 for SVR

Residual mean 0.037

Parameters

- kernel = linear
- C = 0.05
- degree = 2
- epsilon = 0.01

- gamma = scale

Factors (Feature) that have high contributes in predicting target column are as follows

- GDP (High)
- MSACSR (Moderate)
- PERMIT (High)
- MSPUS (High)